Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Editorial

# Updating JPROT's publication standards for large-scale proteomic studies: Towards hypothesis-driven interpretation of predictive biological models

The development in the 1990s of biological mass spectrometry into a robust analytical tool heralded a paradigm shift in biological research. Advances in instrumentation and methodologies have since fueled an expansion of the scope of biological mass spectrometry, from the simple analysis of single proteins to the characterization of highly complex proteomes. The cornerstone of proteomics analysis is the simultaneous identification and quantitation of the protein components of biological systems using increasingly sophisticated and sensitive mass analyzers; its beast, data validation and the interpretation of large-scale experiments. Data validation represents an essential step for ensuring independent repeatability and replication of the experiment by other researchers. However, the toughest aspect of proteomics studies is the interpretation in a meaningful biological context of both limited 2DE gel-based protein maps and large (gel-free, high-throughput) datasets. Around the turn of the new century, proteomics began moving from a purely descriptive mode to quantitative methods. The throughput of proteomics measurements depends on the level of detail desired. Paradoxically, the increased capabilities of mass spectrometry have led to proteomics increasingly functioning as a 'black box' [1]. However, with thousands of genomes sequenced to date, and many other in progress, the goal for the proteomics community now is to show how the 'parts-list' of a complex proteome provides a valuable asset for hypothesis-driven research in biological sciences. The central thesis of this Editorial, developed below, is that validation and interpretation become unified through the formulation of an adequate testable biological hypothesis. A corollary of this statement is that making biological sense of a proteomics experiment requires a paradigm shift in both experimental design and publication policy of biological mass spectrometric data. There is a lack of conformity among proteomics journals about publication standards for proteomics studies, and this is certainly not a trivial matter. Proteomics is applied to investigate a wide range of biological systems, both model (genome sequence available) and non-model organisms, and each study requires a somehow different experimental approach. Although high-quality research on non-model organisms, as well as papers describing significant technical advances in biological mass spectrometry and proteomics, will continue to be welcome in the pages of *Journal of Proteomics*, this journal will implement in the *Instructions for Authors* the requirement for original papers to include a paragraph ("Significance") describing how the reported methodology or proteomics findings significantly advance, respectively, the field and the understanding of the biological process investigated. In addition, authors of studies on model organisms are encouraged to interpret their proteomics data using predictive models. In the medium term this paradigm shift may contribute to change the Linnean taxonomic concept of reporting large collections of proteomics data by a hypothesis-driven approach of elucidating molecular pathways and biological mechanisms.

A key argument for the shift in editorial policy emerges from a reflection on how the information encoded in mass spectra is transformed into a relevant biological output. Mass spectrometry has a long history of use as a standard technique for the structural elucidation of nonpeptidic small molecules with a molecular mass less than 2000 Da [2]. The success of structure elucidation of these small molecules lies in the evaluation of high-resolution and accurate multiple-stage ($MS^n$) mass spectral data based on fragmentation rules rooted in the era of electron ionization mass spectrometry from the 1970s and 1980s (reviewed in [2]), the availability of fragmentation mass spectra libraries from known compounds, and expert computer-aid systems for mass spectral interpretation. Spectral information represents also the Rosetta Stone strategy for deciphering the proteomics experiment. Further, whereas de novo structure elucidation of small molecules by MS is challenging due to the high molecular diversity of structures and fragmentation pathways, including rearrangement reactions, homolytic or heterolytic bond cleavages, hydrogen rearrangements, electron shifts, resonance reactions, and aromatic stabilizations, peptides yield significantly simpler and reliable array of fragment ions [3,4]. Hence, arguably granted, given human or computational skills to decipher it, an MS/MS spectrum containing the complete set of sequence-specific and satellite fragment ions offers sufficient resolution to

1874-3919 
http://dx.doi.org/10.1016/j.jprot.2012.08.022

unambiguously reconstruct the structure of the parent ion. However, assessment of each individual spectrum recorded in high-throughput proteomics measurements is nearly impractical, even for an experienced mass spectrometrist, and an accurate and complete computerized interpretation of fragmentation spectra is far from being a routine application [5]. On the other hand, due to a trade-off between data acquisition speed and spectral quality, most automatically acquired LC-MS/MS spectra, particularly those produced in the limited time of duty cycles of high-throughput or shotgun experiments, do not encode unbroken series of fragment ions; there are gaps and ambiguities and thus decoding the information contained in a partial fragmentation spectrum into a sequence tag requires data reconstruction through probabilistic inference. A major challenge of this approach involves ranking the collection of compatible peptide matches through the assignment of statistical confidence scores that quantify the agreement between observed and predicted data. A significantly high protein score, based on the calculated probability that the observed match between the experimental data and the database sequence is not random, minimizes the number of false positives. However, even the best scoring scheme cannot fully separate the correct and incorrect matches [6]. To circumvent this weakness, a commonly employed strategy to validate probabilistically-assigned matches is the verification of the candidate proteins using independent evidence. As in Schrödinger's cat paradox, orthogonal verification produces the collapse of the assigned probability into a defined state. However, for 2DE and large-scale proteomics workflows the non-observed proteome is often larger than the observed dataset, and data validation per se does not contribute to increase the information content, and thus the biological value, of the observed proteome. On the other hand, the interpretation of the proteomics dataset in a meaningful biological context is critically dependent on the available information. This vicious circle can be opened by verifying predictions based on a model originally generated from the reduced set of experimentally observed data. Each prediction-verification step transfers information from the non-observed to the observed proteomics set, contributing thereby to garner greater amounts of information and thus to a more plausible interpretation of the experiment. The centerpiece of this iterative procedure is the predictive ability of the biological hypothesis. In analogy with particle physics, the hypothesis represents the boson that, through experimental contrasting of its predictions, provides an opportunity to assess the consistency of the proposed model endowing biological meaning to the proteomics data. Hypothesis-driven validation bears the repercussion of reducing the complexity of large-scale datasets and the impact of individual false positives through the recognition of biological trends. A major challenge of model validation is the achievement of an accurate match between model predictions and the experimental data. In this regard, methodologies such as targeted proteome investigation via selected reaction monitoring mass spectrometry [7] may become increasingly important for validating hypothesis-driven model predictions.

Sharing the fruits of research is critically important to the advancement of biological sciences, and there are strong moves towards standards and guidelines for the exchange and publication of proteomics data. The Human Proteome Organization's Proteomics Standards Initiative (http://www.psidev.info/) is very active in this area. From its launch, *Journal of Proteomics* has supported the Proteomics Standards Initiative to develop guidance modules for reporting the use of basic techniques commonly employed in a proteomics workflow [8]. *Journal of Proteomics* also follows with great interest and expectations the development of appropriate databases for information sharing, mining and extracting. A few prototype repositories for proteomics data are starting to get off the ground (i.e. PRIDE, http://www.ebi.ac.uk/pride/), but their structural complexity and lack of appropriate tools still make almost impossible the recovery of valuable knowledge from those sites and thus the possibility of generating hypothesis based on the raw data and associated metadata deposited by other researchers. Authors submitting their work to *J. Proteomics* will be asked to provide access to raw MS data through public databases as they emerge. Meanwhile, data deposition will only be mandatory upon requirement of the editors or reviewers.

Genome databases, such as NCBI (http://www.ncbi.nlm.nih.gov) or GOLD (http://www.genomesonline.org), contain sequence and map data from the whole genomes of several thousands of organisms representing all three domains of life (bacteria, archaea, and eukaryota). Of particular relevance would thus be the development of software to integrate proteomics data gathered on these model organisms into the spectrum of other "omics" data to create metadata structures from which predictive biological models could be generated. Updating *Journal of Proteomics*' guidelines for publishing proteomic studies of model organisms pursues this goal.

## REFERENCES

[1] Smith RD. Trends Biotechnol 2002;20(12, Suppl.):S3-7.
[2] Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanal Rev 2010;2:23-60.
[3] Biemann K. Contributions of mass-spectrometry to peptide and protein-structure. Biomed Environ Mass Spectrom 1988;16:99–111.
[4] Medzihradszky KF. Peptide sequence analysis. Methods Enzymol 2005;402:209-44.
[5] Neuhauser N, Michalski A, Cox J, Mann M. Expert system for computer assisted annotation of MS/MS spectra. Mol Cell Proteomics 2012 doi:10.1074/mcp.M112.020271.
[6] Cottrell JS. Protein identification using MS/MS data. J Proteomics 2011;74:1842-51.
[7] Maiolica A, Jünger MA, Ezkurdia I, Aebersold R. Targeted proteome investigation via selected reaction monitoring mass spectrometry. J Proteomics 2012;75:3495-513.
[8] Calvete JJ. Journal of Proteomics. The first nine months and 6 issues after its Big Bang. J Proteomics 2009;71:573-5.

Juan J. Calvete
*Instituto de Biomedicina de Valencia,*
*Consejo Superior de Investigaciones Científicas,*
*Jaume Roig 11, 46010 Valencia, Spain*
Tel.: +34 96 339 1778; fax: +34 96 369 0800.
*E-mail address:* jcalvete@ibv.csic.es.

29 August 2012